Lecture 15: Bayesian Learning

Fall 2022

Kai-Wei Chang

CS @ UCLA

kw+cm146@kwchang.net

The instructor gratefully acknowledges Dan Roth, Vivek Srikuar, Sriram Sankararaman, Fei Sha, Ameet Talwalkar, Eric Eaton, and Jessica Wu whose slides are heavily used, and the many others who made their course material freely available online.

Announcement

- Hw1 and Midterm are graded

MEDIAN	MAXIMUM	MEAN	STD DEV
92.0	100.0	89.2	11.91

- Hw2 will be due today
- Hw3 will be released today (due 12/5, Mon)
- Final Exam 12/9 11:30am-2:30pm Fri

What you will learn today

- Review Bayesian Theorem
- Maximum a posterior (MAP)
- logistic regression w/ Gaussian prior
- Naïve Bayes

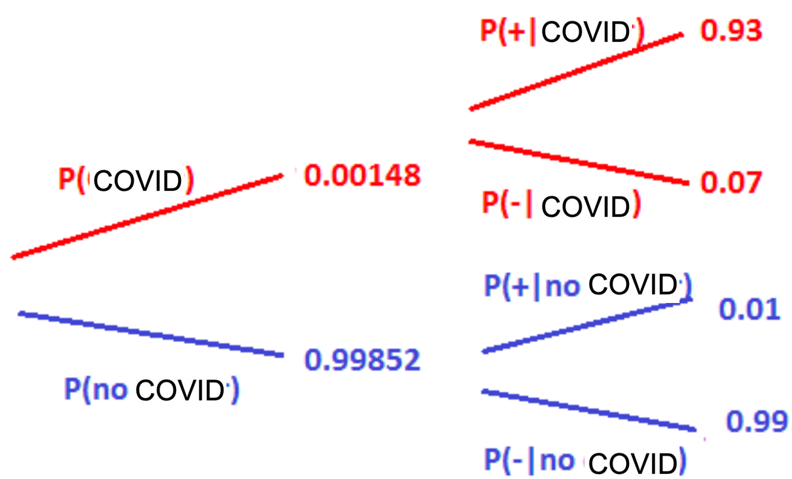
Bayesian Learning

Bayes Theorem Example

- How likely the patient got COVID if the test is positive?

$$\text{P}\,(\text{COVID}\,|+) = \frac{P(\text{COVID } and+)}{P(\text{COVID } and+) + P(no \text{ COVID } and+)} = 0.12$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$



Recap: Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$\big\downarrow \boxed{\text{Short for}}$$

$$\forall x, y \quad P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Prior probability: What is our belief in Y before we see X?

Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

Bayes Theorem

$$\boxed{P(Y|X)} = \frac{\boxed{P(X|Y)}\,\boxed{P(Y)}}{P(X)}$$

Posterior probability: What is the probability of Y given that X is observed?

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

Bayes Theorem

(X, Y) can be (Data, Model), or (Input, Output)

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Posterior probability: What is the probability of Y given that X is observed?

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

$$\forall x, y \quad P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

$$= \frac{P(X = x|Y = y)P(Y=y)}{\sum_{y'} P(X = x|Y = y')P(Y=y')}$$

Probabilistic Learning

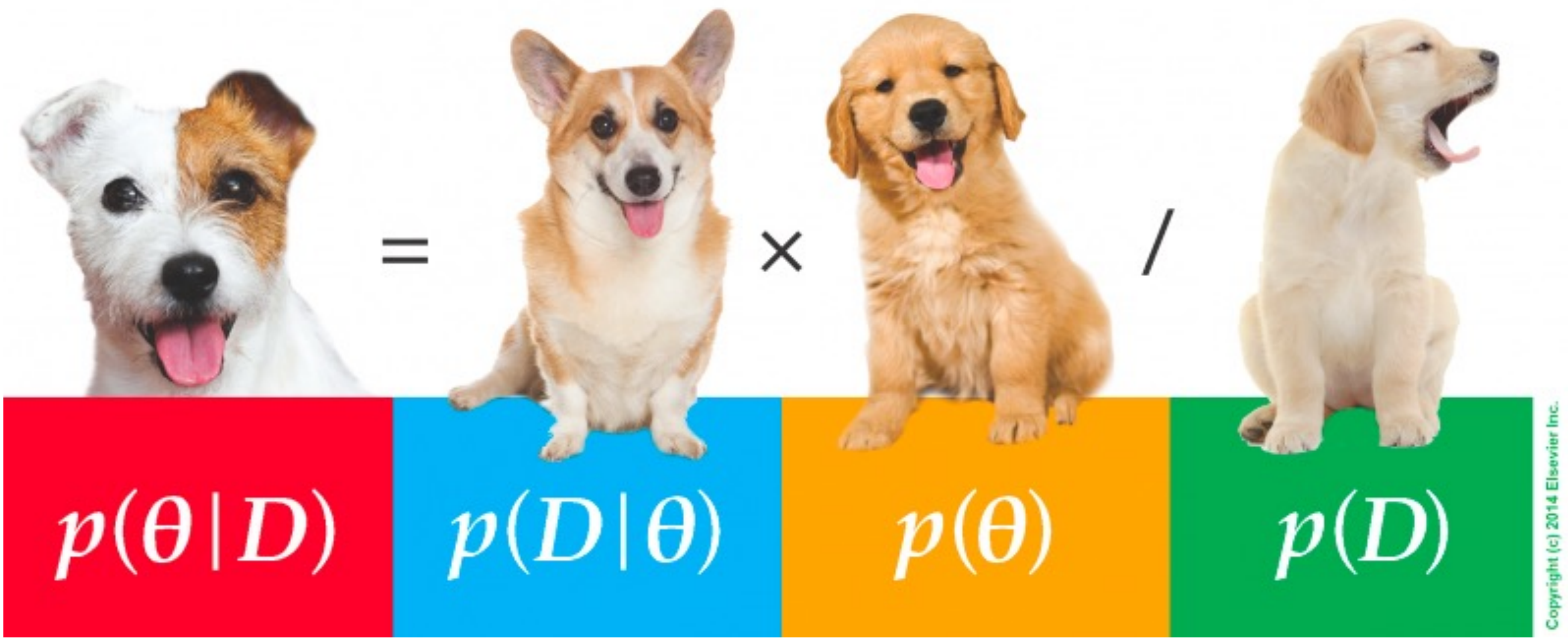
Two different notions of probabilistic learning

- Bayesian Learning: Use of a probabilistic criterion in selecting a hypothesis (P($\Theta|D$))
  - The hypothesis can be deterministic, a Boolean function
  - The criterion for selecting the hypothesis is probabilistic
- Learning probabilistic concepts (P(Y|X))
  - The learned concept is a function $c: X \rightarrow [0,1]$
  - c(x) may be interpreted as the probability that the label 1 is assigned to x

Today’s lecture

- Bayesian Learning
- Maximum a posteriori and maximum likelihood estimation
- Naïve Bayes

Probabilistic models and Bayesian Learning



Bayesian Learning: The basics

- Goal: To find the ***best*** hypothesis from some space H of hypotheses, using the observed data D

- Define best = most probable hypothesis in H

- We assume a probability distribution over the class H

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h

What does *best* mean?

Bayesian learning uses $P(h\,|D)$, the conditional probability of a hypothesis given the data, to define *best*.

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D)$$

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D)$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D)$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Key insight: Both h and D are events.

- D: The event that we observed *this* particular dataset
- h: The event that the hypothesis h is the true hypothesis

<u>So we can apply the Bayes rule here.</u>

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$\boxed{P(h|D)} = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Key insight: Both h and D are events.

- D: The event that we observed *this* particular dataset
- h: The event that the hypothesis h is the true hypothesis

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Prior probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Likelihood: What is the probability that this data point (an example or an entire dataset) is observed, given that the hypothesis is h?

Prior probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

Bayesian Learning

Given a dataset D, we want to find the best hypothesis h
What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Posterior probability: What is the probability that h is the hypothesis, given that the data D is observed?

Likelihood: What is the probability that this data point (an example or an entire dataset) is observed, given that the hypothesis is h?

What is the probability that the data D is observed (independent of any knowledge about the hypothesis)?

Prior probability of h: Background knowledge. What do we expect the hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

Today's lecture

- Bayesian Learning
- Maximum a posteriori and maximum likelihood estimation
- Naïve Bayes

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(h|D) \\
&= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\
&= \arg\max_{h \in H} P(D|h)P(h)
\end{aligned}
$$

Choosing a hypothesis

Given some data, find the most probable hypothesis

❖ The Maximum a Posteriori hypothesis h_{MAP}

P(D) is a constant.

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(h|D) \\
&= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\
&= \arg\max_{h \in H} P(D|h)P(h)
\end{aligned}
$$

Posterior $\propto$ Likelihood $\times$ Prior

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

If we use a uniform distribution for h, then P(h) is constant. Thus, h_MAP = argmax P(D|h) * C = argmax P(D|h), which is the maximum likelihood.

Thus, MAP is equivalent to MLE if h is a uniform distribution.

Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

If we assume that the prior is uniform i.e. $P(h_i) = P(h_j)$, for all h_i, h_j

- We get the Maximum Likelihood

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Often computationally easier to maximize *log likelihood*

Example: Bernoulli trials

The CEO of a startup hires you for your first consulting job

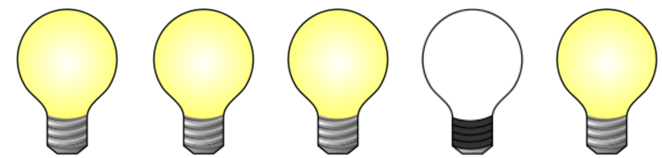
- *CEO*: My company makes light bulbs. I need to know what is the probability they are faulty.
- *You*: Sure. I can help you out. Are they all identical?
- *CEO*: Yes!
- *You*: Excellent. I know how to help. We need to experiment…

Faulty lightbulbs

The experiment:

Try out 100 lightbulbs

80 work, 20 don't



You: The probability is P(failure) = 0.2

CEO: But how do you know?

You: Because…

Bernoulli trials

- P(success) = p, P(failure) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

Bernoulli trials

- P(success) = p, P(failure) = 1 – p

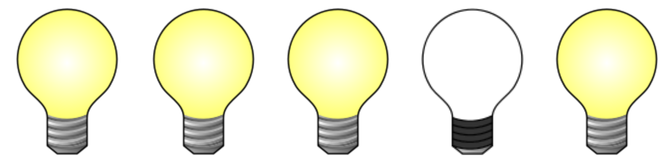
- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

Bernoulli trials

- P(success) = p, P(failure) = 1 – p



- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

- The most likely value of p for this observation is?

Bernoulli trials

- P(success) = p, P(failure) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

- The most likely value of p for this observation is?

$$\underset{p}{\operatorname{argmax}} P(D|p) = \underset{p}{\operatorname{argmax}} \binom{100}{80} p^{80}(1-p)^{20}$$

The "learning" algorithm

Say you have *a* Work and b Not-Work

$$\begin{aligned} p_{best} &= \operatorname*{argmax}_p P(D|h) \\ &= \operatorname*{argmax}_p \log P(D\,|h) \\ &= \operatorname*{argmax}_p \log\left(\binom{a+b}{a} p^a (1-p)^b\right) \\ &= \operatorname*{argmax}_p a\log p + b\log(1-p) \end{aligned}$$

Calculus 101: Set the derivative to zero

$$P_{best} = a/(a+b)$$

The "learning" algorithm

Say you have *a* Work and b Not-Work

$$
\begin{aligned}
p_{best} &= \underset{p}{\operatorname{argmax}} P(D|h) \\
&= \underset{p}{\operatorname{argmax}} \log P(D\,|h) \\
&= \underset{p}{\operatorname{argmax}} \log\left(\binom{a+b}{a} p^a (1-p)^b\right) \\
&= \underset{p}{\operatorname{argmax}}\, a \log p + b \log(1-p)
\end{aligned}
$$

Log likelihood

Calculus 101: Set the derivative to zero

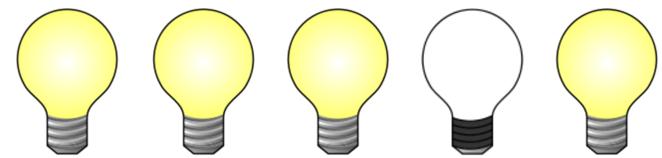
$$P_{best} = a/(a + b)$$

Faulty lightbulbs

The experiment:

Try out 100 lightbulbs

80 work, 20 don't



You: The probability is P(failure) = 0.2

CEO: But how do you know?

You: Because…

CEO: Okay, but can you incorporate some results from our prior tests?

MAP estimation

Given some data, find the most probable hypothesis

- The Maximum a Posteriori hypothesis h_{MAP}

$$h_{MAP} = \arg\max_{h \in H} P(D|h) \boxed{P(h)}$$

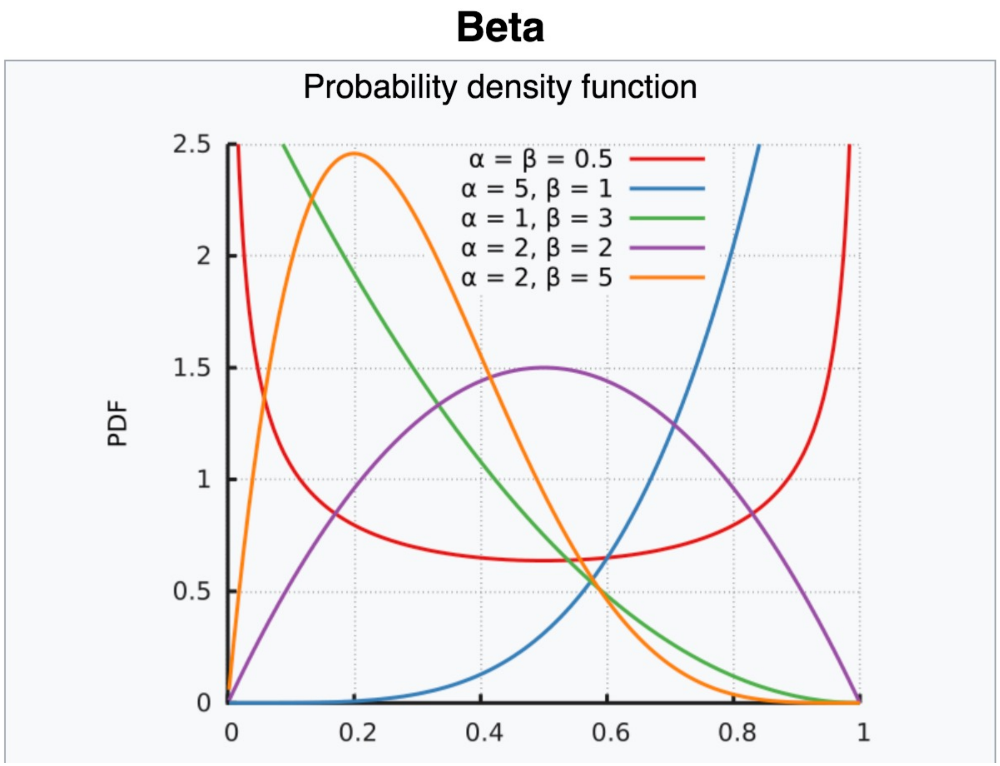
If we assume that the prior is uniform i.e. $P(h_i) = P(h_j)$, for all h_i, h_j

- Simplify this to get the Maximum Likelihood hypothesis

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Often computationally easier to maximize *log likelihood*

Beta distribution



$$
\begin{aligned}
f(x;\alpha,\beta) &= \mathrm{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$

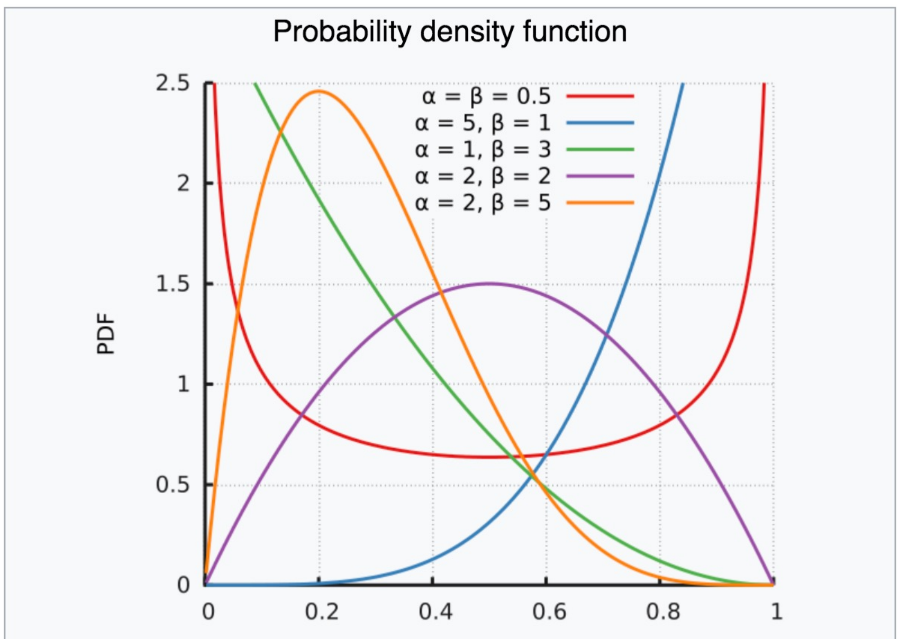
$\alpha > 0, \beta > 0$

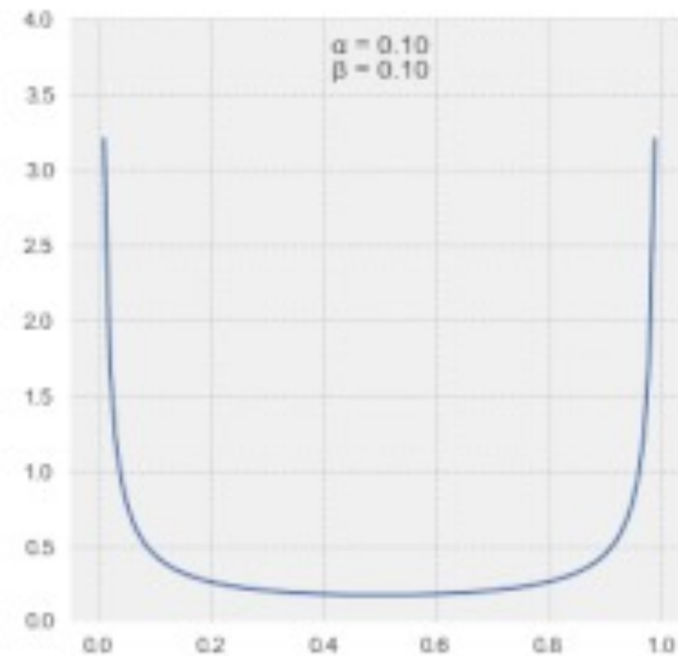
https://en.wikipedia.org/wiki/Beta_distribution

Prior distribution

- The boss has a prior belief of the distribution of faulty lightbulb

Beta





For aloha = beta = large, then the beta distribution is symmetric, and the best p wold be in the middle.

MAP for Bernoulli trials

p is the parameter for hypothesis

- P(success) = p, P(failure) = 1 – p

- Each trial is i.i.d
  - Independent and identically distributed

- You have seen D = {80 work, 20 don't}

$$P(D|p) = \binom{100}{80} p^{80}(1-p)^{20}$$

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

MAP estimation

Assuming h is distributed according to Beta distribution

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

$$P(D|p) = \binom{a+b}{a} p^a (1-p)^b$$

$$P(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$p_{best} = \underset{p}{\operatorname{argmax}}\, P(D|h)\, P(h)$$

$$= \underset{p}{\operatorname{argmax}} \log P(D\,|h) + \log P(h)$$

$$= \underset{p}{\operatorname{argmax}} \log\left(\frac{\binom{a+b}{a}}{B(\alpha,\beta)} p^a (1-p)^b p^{\alpha-1}(1-p)^{\beta-1}\right)$$

$$= \underset{p}{\operatorname{argmax}}\ (a+\alpha-1)\log p + (b+\beta-1)\log(1-p)$$

MAP v.s. MLE

- MLE:

$$\operatorname*{argmax}_p a \log p + b \log(1-p)$$

$$\Rightarrow p_{best} = \frac{a}{a+b}$$

- MAP (w/ Beta distribution as prior)

$$\operatorname*{argmax}_p (a + \alpha - 1) \log p + (b + \beta - 1) \log(1-p)$$

$$\Rightarrow p_{best} = \frac{a + \alpha - 1}{a + b + \alpha + \beta - 2}$$

MAP v.s. MLE

- MAP

  $$\underset{p}{\operatorname{argmax}}\ (a+\alpha-1)\log p+(b+\beta-1)\log(1-p)$$

  $$\Rightarrow p_{best}=\frac{a+\alpha-1}{a+b+\alpha+\beta-2}$$

- Let $\alpha=100,\ \ \beta=10$
- $a=10, b=20 \Rightarrow p_{best} \approx 0.79$
- $a=1000, b=2000 \Rightarrow p_{best} \approx 0.36$
- $a=100{,}000, b=200{,}000 \Rightarrow p_{best} \approx 0.33$

As a and b go up, p_best ~ a/(a + b), which is the same as the MLE

MAP for logistic regression

Let's review MLE for logistic regression

- Training data
  - $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, m examples
- What we want
  - Find a **w** such that P(S | **w**) is maximized
  - We know that our examples are drawn independently and are identically distributed (i.i.d)
  - How do we proceed?

Maximum likelihood estimation

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

The usual trick: Convert products to sums by taking log

Recall that this works only because log is an increasing function and the maximizer will not change

Maximum likelihood estimation

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

$\boxed{\text{Equivalent to solving}}$

$$P(y|\mathbf{w}, \mathbf{x}) = \frac{1}{1 + \exp(-\mathrm{y_i}(\mathbf{w}^T \mathbf{x}_i + b))}$$

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i(\mathbf{w}^T \mathbf{x}_i + b)))$$

Maximum likelihood estimation

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

The goal: Maximum likelihood training of a discriminative probabilistic classifier under the logistic model for the posterior distribution.

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b)))$$

Equivalent to: Training a linear classifier by minimizing the ***logistic loss***.

Maximum a posteriori estimation

We could also add a prior on the weights

Suppose each weight in the weight vector is drawn independently from the normal distribution with zero mean and standard deviation σ

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Adding this probabilty distribution helps regularize the logistic regression and avoid ovdrfitting of data. Sigma helps balance this regularization term.

MAP estimation for logistic regression

$$\underset{\mathbf{w}}{\operatorname{argmax}} P(S|\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$P(y|\mathbf{w}, \mathbf{x}) = \frac{1}{1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))}$$

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b)))$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\underset{\mathbf{w}}{\operatorname{argmax}} P(S|\mathbf{w})P(\mathbf{w})$$

Take log to simplify

$$\underset{\mathbf{w}}{\operatorname{argmax}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

MAP estimation for logistic regression

$$\underset{\mathbf{w}}{\text{argmax}}\, P(S|\mathbf{w}) = \underset{\mathbf{w}}{\text{argmax}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$P(y|\mathbf{w}, \mathbf{x}) = \frac{1}{1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))}$$

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b)))$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\underset{\mathbf{w}}{\text{argmax}}\, P(S|\mathbf{w})P(\mathbf{w})$$

Take log to simplify

$$\underset{\mathbf{w}}{\text{argmax}}\, \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

This is the log-likelihood

We have already expanded out the first term.

$$\sum_{i}^{m} -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b)))$$

MAP estimation for logistic regression

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_i \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$P(y|\mathbf{w}, \mathbf{x}) = \frac{1}{1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))}$$

$$\max_{\mathbf{w}} \sum_i^m -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b)))$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\operatorname*{argmax}_{\mathbf{w}} P(S|\mathbf{w})P(\mathbf{w})$$

Take log to simplify

$$\operatorname*{argmax}_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

Expand the log prior

$$\sum_i^m -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))) + \sum_{j=1}^{d} \frac{-w_i^2}{\sigma^2} + constants$$

MAP estimation for logistic regression

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i, \mathbf{w})$$

$$\max_{\mathbf{w}} \sum_{i}^{m} \log P(y_i|\mathbf{x}_i, \mathbf{w})$$

Equivalent to solving

$$P(y|\mathbf{w}, \mathbf{x}) = \frac{1}{1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))}$$

$$\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b)))$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_i) = \prod_{j=1}^{d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{\sigma^2}\right)$$

Learning by solving

$$\arg\max_{\mathbf{w}} P(S|\mathbf{w})P(\mathbf{w})$$

Take log to simplify

$$\arg\max_{\mathbf{w}} \log P(S|\mathbf{w}) + \log P(\mathbf{w})$$

$$\arg\max_{\mathbf{w}} \sum_{i}^{m} -\log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))) - \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w}$$

Maximizing a negative function is the same as minimizing the function

$$\arg\min_{\mathbf{w}} \sum_{i}^{m} \log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))) + \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w}$$

Learning a logistic regression classifier

Learning a logistic regression classifier is equivalent to solving

As sigma goes up, guassian distribution approaches uniform distribution, and this term goes down to 0. Note that uniform prior MAP is equivalent to MLE

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i}^{m} \log(1 + \exp(-y_i(\mathbf{w}^T\mathbf{x}_i + b))) + \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w}$$

LOGISTIC LOSS

l2-regularization (after adding a guassian prior)

For logistic regression, we add this term due to guassian prior (P(w)). For SVM, this term gives you the largest margin.

Naïve Bayes

Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Posterior probability: What is the probability of Y given that X is observed?

Likelihood: What is the likelihood of observing X given a specific Y?

Prior probability: What is our belief in Y before we see X?

Probabilistic Learning

Two different notions of probabilistic learning

- Bayesian Learning: Use of a probabilistic criterion in selecting a hypothesis ($P(\Theta|D)$)
  - The hypothesis can be deterministic, a Boolean function
  - The criterion for selecting the hypothesis is probabilistic
- Learning probabilistic concepts ($P(Y|X)$)
  - The learned concept is a function $c: X \rightarrow [0,1]$
  - $c(x)$ may be interpreted as the probability that the label 1 is assigned to x

MAP prediction

log-regression: P(y|x) = sig(w^T x + b)

Let's be use the Bayes rule for predicting y given an input **x**

$$P(Y = y|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

Posterior probability of label being y for this input x

Best assignment of labels to give largest probability P(Y | X)

MAP prediction

Let's be use the Bayes rule for predicting y given an input **x**

$$P(Y = y|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

P(x) is a constant

Predict y for the input **x** using

$$\arg\max_{y} \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

For a given x, find y that maximizes P(y|x)

MAP prediction

Let's be use the Bayes rule for predicting y given an input **x**

$$P(Y = y|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

Predict y for the input **x** using

$$\arg\max_{y} P(X = \mathbf{x}|Y = y)P(Y = y)$$

MAP prediction

Don't confuse with *MAP learning*: finds hypothesis by

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

Let's be use the Bayes rule for predicting y given an input **x**

$$P(Y = y|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = y)P(Y = y)}{P(X = \mathbf{x})}$$

Predict y for the input **x** using

$$\arg\max_{y} P(X = \mathbf{x}|Y = y)P(Y = y)$$

MAP prediction

Predict y for the input **x** using

$$\arg\max_y P(X = \mathbf{x} | Y = y) P(Y = y)$$

Likelihood of observing this input **x** when the label is y

Prior probability of the label being y

All we need are these two sets of probabilities

Example: Tennis

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

Find T that maximizes P(T | input) = P(T | Temp = Hot and Wind = Weak)

Likelihood

Temperature	Wind	P(T, W | Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W | Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Example: Tennis

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Without any other information, what is the prior probability that I should play tennis?

Example: Tennis

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Without any other information, what is the prior probability that I should play tennis?

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

On days that I do play tennis, what is the probability that the temperature is T and the wind is W?

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

On days that I don't play tennis, what is the probability that the temperature is T and the wind is W?

Example: Tennis

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

Example: Tennis

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

argmax_y P(H, W | play?) P (play?)

Example: Tennis

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W | Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W | Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

$\text{argmax}_y\ P(H, W \mid \text{play?})\ P(\text{play?})$

$P(H, W \mid \text{Yes})\ P(\text{Yes}) = 0.4 \times 0.3 = 0.12$

$P(H, W \mid \text{No})\ P(\text{No}) = 0.1 \times 0.7 = 0.07$

MAP prediction = Yes

How hard is it to learn probabilistic models?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook: S(unny), O(vercast), R(ainy)

Temperature: H(ot), M(edium), C(ool)

Humidity: H(igh), N(ormal), L(ow)

Wind: S(trong), W(eak)

How hard is it to learn probabilistic models?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook: S(unny), O(vercast), R(ainy)

Te

Hu

N(ormal), L(ow)

Wind: S(trong), W(eak)

We need to learn

1. The prior P(Play?)
2. The likelihoods P(X | Play?)

How hard is it to learn probabilistic models?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Prior P(play?)

- A single number (Why only one?)

How hard is it to learn probabilistic models?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Prior P(play?)

- A single number (Why only one?)

Likelihood P(**X** | Play?)

- There are 4 features
- For each value of Play? (+/-), we need a value for each possible assignment: $P(x_1, x_2, x_3, x_4 \mid \text{Play?})$

How hard is it to learn probabilistic models?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-
	3	3	3	2	

Values for this feature

Prior P(play?)

- A single number (Why only one?)

Likelihood P(**X** | Play?)

- There are 4 features
- For each value of Play? (+/-), we need a value for each possible assignment: $P(x_1, x_2, x_3, x_4 \mid \text{Play?})$

How hard is it to learn probabilistic models?

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-
	3	3	3	2	

Values for this feature

Prior P(play?)

- A single number (Why only one?)

Likelihood P(**X** | Play?)

- There are 4 features
- For each value of Play? (+/-), we need a value for each possible assignment: $P(x_1, x_2, x_3, x_4 \mid \text{Play?})$
- $\boxed{(3 \cdot 3 \cdot 3 \cdot 2 \ - 1)}$ parameters in each case

a lot of parameters!!

One for each assignment

Need a lot of data to estimate these many numbers!

How hard is it to learn probabilistic models?

Prior P(Y)

- If there are k labels, then k – 1 parameters

Likelihood P(**X** | Y)

  - We need a value for each possible $P(x_1, x_2, \cdots, x_d \mid y)$ for each y
  - Need a lot of parameters!!

Need a lot of data to estimate these many numbers!

High model complexity

If there is very limited data, high variance in the parameters

How can we deal with this?

Answer: Make independence assumptions